INFORMATION AND CONTROL 19, 174–179 (1971)

An Information Measure of Association in Contingency Tables

M. A. HAMDAN* AND CHRIS P. TSOKOS

Virginia Polytechnic Institute, and State University, Blacksburg, Virginia 24061

Linfoot (1957) introduced an informational measure r_I of correlation between two random variables X and Y. The measure r_I is based on the information gain r_0 in knowing that X and Y are mutually dependent with a given bivariate density function as compared with the original knowledge that X and Y are statistically independent. In the present paper, an asymptotic form of the information measure r_I, denoted by $\tilde{r}_I$, is derived in terms of Pearson's (1904) chi-square for contingency tables. Hence $\tilde{r}_I$ is suggested as an information measure of association in contingency tables. On comparing $\tilde{r}_I$ with Pearson's classical coefficient of contingency P, it is found that $\tilde{r}_I \geqslant P$. This is a desirable property of $\tilde{r}_I$, since Lancaster and Hamdan (1964) demonstrated that P underestimates the product-moment correlation coefficient in contingency tables with broad categories. The asymptotic variance of $\tilde{r}_I$ is derived and compared with the asymptotic variance of P.

1. INTRODUCTION

Given two random variables X and Y with bivariate density function $p(x, y)$ and marginal density functions $p_1(x)$ and $p_2(y)$, Linfoot (1957) defined

$$r_0 = \int\int p(x, y) \ln \left[\frac{p(x, y)}{p_1(x) p_2(y)} \right] dx\, dy \tag{1.1}$$

as the information gain in knowing the bivariate density $p(x, y)$ as compared to the knowledge that X and Y are statistically independent. If X and Y have a bivariate normal distribution with correlation coefficient ρ, then $r_0 = (-\frac{1}{2}) \ln(1 - \rho^2)$, which led Linfoot (1957) to suggest an informational measure r_I of ρ defined by

$$r_I = [1 - \exp(-2r_0)]^{1/2}. \tag{1.2}$$

* On leave from The American University of Beirut.

For discrete random variables X and Y with bivariate probabilities p_{ij} $(i = 1, 2,..., s; j = 1, 2,..., t)$ and marginal probabilities $p_{i.} = \sum_j p_{ij}$ and $p_{.j} = \sum_i p_{ij}$, r_0 takes the form

$$r_0 = \sum_i \sum_j p_{ij} \ln\left[\frac{p_{ij}}{p_{i.}p_{.j}}\right]. \tag{1.3}$$

It should be mentioned here that r_0 is equivalent to Kullback's (1959) distance-like measure of information discrimination between the correlated bivariate distribution $p(x, y)$ (or p_{ij}) and the distribution under independence, $p_1(x)p_2(y)$ (or $p_{i.}p_{.j}$). Furthermore, r_0 may be defined as the expected value of $\ln[p(x, y)/(p_1(x)p_2(y))]$ with respect to the bivariate density function $p(x, y)$ (cf., Good (1956)). The measure r_0 may also be expressed in terms of the Radon–Nikodym derivative $\Omega(x, y)$ as an expectation with respect to the independence distribution $p_1(x)p_2(y)$,

$$r_0 = \int\int \Omega(x, y) \ln[\Omega(x, y)] p_1(x)p_2(y)\, dx\, dy \tag{1.4}$$

where $\Omega(x, y) = p(x, y)/p_1(x)p_2(y)$.

2. The Asymptotic Form of r_0 for Contingency Tables

Consider n observations on a bivariate random variable (X, Y) classified in the form of an $s \times t$ contingency table. Let n_{ij}, $n_{i.}$, $n_{.j}$ $(i = 1, 2,..., s; j = 1, 2,..., t)$ be the observed frequencies in the (i, j) cell, i-th row and j-th column, respectively. Here $\sum_i \sum_j n_{ij} = n$, $\sum_j n_{ij} = n_{i.}$ and $\sum_i n_{ij} = n_{.j}$. We shall denote by p_{ij}, $p_{i.}$, and $p_{.j}$ the probabilities that an observation falls in the (i, j) cell, i-th row, and j-th column, respectively. The maximum likelihood estimates of these multinomial probabilities are given by Kendall & Stuart (1967, p. 548) and Wilks (1962, p. 425):

$$\hat{p}_{ij} = n_{ij}/n, \qquad \hat{p}_{i.} = n_{i.}/n, \qquad \hat{p}_{.j} = n_{.j}/n. \tag{2.1}$$

Hence, the corresponding estimate of r_0 for the $s \times t$ contingency table is

$$\hat{r}_0 = \sum_i \sum_j (n_{ij}/n) \ln(n_{ij}/e_{ij}), \tag{2.2}$$

where $e_{ij} = n_{i.}n_{.j}/n$ is the expected frequency in the (i, j) cell under the assumption of independence, i.e., a bivariate density function of the form

$(p_{i.}p_{.j})$. Writing $D_{ij} = n_{ij} - e_{ij}$, so that $n_{ij} = e_{ij}(1 + D_{ij}/e_{ij})$, formula (2.2) becomes

$$\hat{r}_0 = (1/n)\sum_i \sum_j e_{ij}(1 + D_{ij}/e_{ij}) \ln(1 + D_{ij}/e_{ij}). \tag{2.3}$$

An asymptotic form of the estimator $\hat{r}_0$ is of significant importance from the theoretical and practical points of view. Such a form may be derived by expanding $\ln(1 + D_{ij}/e_{ij})$ up to D_{ij}^2. Thus we get an asymptotic form $\tilde{r}_0$ given by

$$\begin{aligned}\tilde{r}_0 &= \frac{1}{n}\sum_i \sum_j e_{ij}\left(1 + \frac{D_{ij}}{e_{ij}}\right)\left[\frac{D_{ij}}{e_{ij}} - \frac{1}{2}\left(\frac{D_{ij}}{e_{ij}}\right)^2\right] \\ &= \frac{1}{n}\sum_i \sum_j D_{ij} + \frac{1}{2n}\sum_i \sum_j D_{ij}^2/e_{ij}\,. \end{aligned} \tag{2.4}$$

The first term on the right side of (2.4) is zero since $\sum_i \sum_j n_{ij} = n$ and $\sum_i \sum_j e_{ij} = n$. Therefore,

$$\tilde{r}_0 \approx \chi^2/(2n) = \phi^2/2, \tag{2.5}$$

where χ^2 is the usual Pearson chi-square for the $s \times t$ contingency table, namely,

$$\chi^2 = \sum_i \sum_j (n_{ij} - e_{ij})^2/e_{ij} = n\left[\sum_i \sum_j \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right], \tag{2.6}$$

and $\phi^2 = \chi^2/n$ is Pearson's (1904) mean square contingency.

Formula (2.5) provides a useful asymptotic estimate of r_0 based on the information available in the $s \times t$ contingency table, which is a random sample of size n from the bivariate distribution of X and Y. The calculation of $\tilde{r}_0$ presents no difficulty since it only requires the calculation of Pearson's χ^2, as given by formula (2.6).

3. An Information Measure of Correlation in Contingency Tables Derived from the Bivariate Normal Distribution

If the bivariate distribution of the random variable (X, Y) underlying the contingency table is bivariate normal with correlation coefficient ρ, we may derive an asymptotic form of Linfoot's (1957) informational measure of

correlation r_I. This asymptotic form of r_I, denoted by $\tilde{r}_I$, is obtained by substituting $\tilde{r}_0$ for r_0 in formula (1.2), thus getting

$$\tilde{r}_I = [1 - \exp(-2\tilde{r}_0)]^{1/2} = [1 - \exp(-\phi^2)]^{1/2}. \tag{3.1}$$

Obviously, $0 \leqslant \tilde{r}_I \leqslant 1$ with $\tilde{r}_I = 0$ if and only if $\phi^2 = 0$ or $\chi^2 = 0$ or $n_{ij} = e_{ij}$ for all i and j, i.e., X and Y are independent. However, $\tilde{r}_I$ does not attain 1 in the case of complete association, i.e., when all nonzero frequencies lie on a longest diagonal of the $s \times t$ table. In this case of complete association, the maximum value of χ^2 is $[n \min(s - 1, t - 1)]$; cf., Kendall and Stuart (1967, p. 557). Thus, we have

$$\max \tilde{r}_I = [1 - \exp\{-\min(s - 1, t - 1)\}]^{1/2}. \tag{3.2}$$

It is interesting to compare $\tilde{r}_I$ with Pearson's coefficient of contingency (Kendall and Stuart, 1967, p. 557):

$$P = \left(\frac{\phi^2}{1 + \phi^2}\right)^{1/2}. \tag{3.3}$$

In fact, since $\phi^2 \geqslant 0$, we have

$$1 - e^{-\phi^2} \geqslant 1 - \frac{1}{1 + \phi^2} = \frac{\phi^2}{1 + \phi^2}, \tag{3.4}$$

so that $\tilde{r}_I \geqslant P$, with the equality holding only if $\phi^2 = 0$, that is, X and Y are independent. The fact that $\tilde{r}_I \geqslant P$ is a desirable property for two reasons. First, the coefficient of contingency P also has the deficiency of not attaining unity in the case of complete association in the contingency table. In fact,

$$\max P = \left[\frac{\min(s - 1, t - 1)}{1 + \min(s - 1, t - 1)}\right]^{1/2}. \tag{3.5}$$

From Eqs. (3.2) and (3.4) together with (3.5), we have

$$\max \tilde{r}_I \geqslant \max P, \tag{3.6}$$

i.e., in the case of complete association, $\tilde{r}_I$ is closer to unity than P. Secondly, it was demonstrated by Lancaster and Hamdan (1964) (see also Kendall and Stuart, 1967, p. 561) that P underestimates the correlation coefficient for contingency tables with broad categories. This corresponds to a bivariate sample grouped widely in a contingency table with relatively small dimensions; so that the contingency table involves a considerable loss of information as compared with the original ungrouped data.

The asymptotic variances of P and $\tilde{r}_I$ may now be compared. The asymptotic variance of P (Kendall and Stuart, 1967, p. 560) is

$$\text{var } P \approx \left(\frac{\partial P}{\partial \phi^2}\right)^2 \text{var } \phi^2 = \frac{1}{4\phi^2(1+\phi^2)^3} \text{var } \phi^2. \tag{3.7}$$

Similarly,

$$\begin{aligned} \text{var } \tilde{r}_I &\approx \left(\frac{\partial \tilde{r}_I}{\partial \phi^2}\right)^2 \text{var } \phi^2 \\ &= \tfrac{1}{4}[e^{\phi^2}(e^{\phi^2}-1)]^{-1} \text{var } \phi^2. \end{aligned} \tag{3.8}$$

It should be noted that (3.7) and (3.8) are not valid when the two variables are independent or $\phi^2 = 0$. The variance of ϕ^2 is $1/n^2$ var χ^2; the variance of χ^2 under nonindependence was derived by Pearson (1915). The ratio of (3.8) to (3.7) is

$$\frac{\text{var } \tilde{r}_I}{\text{var } P} \approx \frac{\phi^2(1+\phi^2)^3}{e^{\phi^2}(e^{\phi^2}-1)}. \tag{3.9}$$

A numerical study of formula (3.9), to one decimal place for values of ϕ^2 shows that the ratio is greater than unity for $0 < \phi^2 < 2.2$, and less than unity for $\phi^2 \geqslant 2.2$.

4. Concluding Remarks

For $s \times t$ contingency tables in general, the asymptotic form $\tilde{r}_0$ of Linfoot's (1957) measure of information discrimination is suggested as a measure of the degree of association in the table. For contingency tables with an underlying bivariate normal distribution, $\tilde{r}_I$ is suggested as an information measure of the correlation coefficient ρ. Like P, Pearson's coefficient of contingency, $\tilde{r}_I$ does not attain unity in the case of complete association. However, $\tilde{r}_I$ is always larger than P and hence closer to unity for complete association. Besides $\tilde{r}_I$ is preferable to P In the case of contingency tables with broad classes, where P is known to underestimate ρ. The asymptotic variance of $\tilde{r}_I$ is smaller or larger than that of P depending on whether $\phi^2 \geqslant 2.2$ or $0 < \phi^2 < 2.2$, respectively.

Received: August 5, 1970; Revised: March 22, 1971

REFERENCES

GOOD, I. J. (1956), Some Terminology and Notation in Information Theory, Proceedings of the Institute of Electrical Engineers **103**, 200–204.

KENDALL, M. G., AND STUART, A. (1967), "The Advanced Theory of Statistics," Vol. 2, Charles Griffin, London.

KULLBACK, S. (1959), "Information Theory and Statistics," John Wiley and Sons, New York.

LANCASTER, H. O., AND HAMDAN, M. A. (1964), Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characters, *Psychometrika* **29**, 383–391.

LINFOOT, E. H. (1957), An informational measure of correlation, *Information and Control* **1**, 85–89.

PEARSON, K. (1904), "On the Theory of Contingency and its Relation to Association and Normal Correlation," Drapers' Co. Memoirs, Biometric Series 1, London,

PEARSON, K. (1915), On the probable error of a coefficient of mean square contingency, *Biometrika* **10**, 570–573.

WILKS, S. S. (1962), "Mathematical Statistics," John Wiley and Sons, New York.

Printed in Belgium